Class Project 1

Bakiri Ayman, Ben Mohamed Nizar, Chahed Ouazzani Adam
CS-433 : Machine learning, EPFL

***Abstract*—In this study, we developed predictive models for heart disease risk assessment using a medical dataset with over 300,000 samples and 321 features. Given the high dimensionality and severe class imbalance, we explored various preprocessing methods, feature engineering, and regularized machine learning techniques. Ridge regression emerged as the most effective model, with an F1 score of 0.4328. By implementing optimal hyperparameter tuning and a threshold adjustment function, we improved model performance significantly, emphasizing the importance of both feature selection and model calibration in high-stakes medical applications.**

I. Introduction

This project focuses on analyzing a medical dataset to build predictive models for health-related outcomes. Using various machine learning techniques, we explore the dataset, preprocess the data, and test multiple algorithms. Our main goal is to evaluate model performance and determine the best approach for predicting health outcomes, prioritizing accuracy and F1 score.

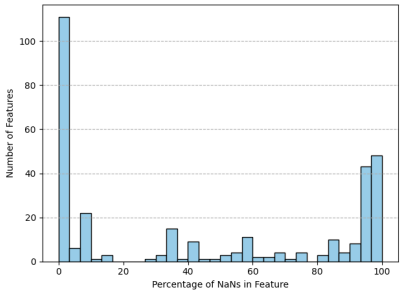
II. Dataset Exploration

The dataset includes 328,135 training samples and 109,379 test samples, each with 321 features covering health behaviors and demographics. We found a substantial class imbalance, with 299,160 samples labeled -1 (majority) and 28,975 labeled 1 (minority). The dataset's high dimensionality and missing values led us to employ feature selection techniques.

III. DataSet Pre-Processing

For such a medical dataset, the Pre-processing step was crucial for enhancing our ML models performance. Given the complexity of medical data, we used various techniques to clean and prepare the dataset. These steps ensure that the model can learn meaningful patterns while minimizing noise and handling class imbalance

Dropping Columns with High NaN Ratios



Distribution of Missing Values Across Features

The plot shows that over 100 features contain 90 percent missing values. To reduce noise and dimensionality, we removed columns with more than 60 percent missing values, ensuring the dataset retains only relevant and informative features.

Fill Missing Values

To handle missing values effectively, we used a method that differentiates between discrete and continuous features. For discrete features, missing values were filled with a neutral category value, which serves as a placeholder without introducing biases. For continuous features, missing values were filled with the mean of the feature.

Features Engineering

Polynomial features up to degree 3 were created to capture non-linear relationships between features and the target variable.

Removing Highly Correlated Features

Features with a correlation above 0.8 were removed to reduce redundancy, ensuring a consistent selection between training and test sets.

Removing Low-Variance Features

Features with variance below 0.01 were removed, as they contribute little information to the model, retaining only informative features.

Feature Normalization

Each feature was standardized by subtracting the training set mean and dividing by the standard deviation, resulting in a mean of 0 and a standard deviation of 1. This standardization process ensures that all features are on a comparable scale, preventing any single feature from disproportionately influencing the model's learning.

Bias Term Addition

A bias term was added to the feature matrix for both logistic and ridge regression models. This term enables the models to fit an intercept, which enhances their flexibility in capturing the underlying data distribution.

Class Balancing

We observed a significant class imbalance in the dataset, with 299,160 negative samples and only 28,975 positive samples, resulting in an approximate 1 to 10 ratio. To address this, we upsampled the minority class by a factor of 2. This approach involved duplicating a fraction of the minority samples, with slight noise added to each duplicate for variability. By balancing the classes in this way, we aim to mitigate the

model's tendency to overfit to the majority class and improve the F1 score.

IV. Evalutation Metrics

We used accuracy and F1 score to evaluate our model. While accuracy measures the ratio of correct predictions to total instances, it can be misleading in our imbalanced medical dataset. For example, predicting only the majority class yields a high accuracy of 91 percent, despite poor minority class performance. The F1 score, balancing precision and recall, provides a clearer assessment for imbalanced classes, making it our primary evaluation metric.

V. Models

We experimented with various models, preprocessing combinations, and hyperparameter tuning. We began with logistic regression, which was well-suited to the classification task. Then, we moved to ridge regression, leveraging its closed-form solution and polynomial feature expansion, which ultimately yielded the best results.

A. *Logistic regression*

Motivation :

Logistic regression addresses the classification of heart conditions based on health-related features. This model is critical due to its ability to provide probabilistic outputs and interpretability, which are essential in medical applications.

Regularization Effect :

Regularization is applied to prevent overfitting, particularly given the dataset's high dimensionality. The model is optimized using gradient descent, allowing for efficient training on the large sample size.

Gamma \ Lambda	0.0001	0.001	0.01
0.001	0.3324	0.3324	0.3324
0.01	0.4044	0.4043	0.4042
0.1	0.4046	0.4046	0.4064
0.5	0.4105	0.4099	0.4065
1	0.4103	0.4103	0.4065

Validation Set F1 Scores for different combinations of gamma and lambdas

We see that with a lambda of 0.0001 and gamma of 0.5, logistic regression achieved the highest f1 score of 0.4105.

B. *Ridge Regression*

Motivation: Ridge regression addresses heart condition classification by applying regularization to manage model complexity. It is well-suited to high-dimensional medical data, reducing the impact of less informative features and stabilizing predictions.

Regularization Effect: Regularization, controlled by the lambda parameter, limits large weights, helping prevent overfitting and improving generalization. Ridge regression, optimized with gradient descent and a tuned learning rate, maintains predictive accuracy while handling dataset variance.

Parameter Tuning: We optimized ridge regression by selecting the best lambda and classification threshold with two different methods, and iterating on preprocessing steps to observe and refine performance improvements.

Preprocessing Steps	Best Accuracy and F1 Score
• Remove NaN features • Fill with mean • Remove low-variance features • Remove highly correlated features	**Train:** Accuracy = 0.8656, F1 Score = 0.4044 **Validation:** Accuracy = 0.8646, F1 Score = 0.4119
• Add normalization • Add bias term	**Train:** Accuracy = 0.8682, F1 Score = 0.4040 **Validation:** Accuracy = 0.8670, F1 Score = 0.4120
• Upsample minority class	**Train:** Accuracy = 0.8157, F1 Score = 0.5398 **Validation:** Accuracy = 0.8736, F1 Score = 0.4145
• Feature expansion (degree 3) • Retain correlated features • Optimal lambda: 0.000022, Optimal threshold: -0.1191	**Train:** Accuracy = 0.8238, F1 score = 0.5664 **Validation:** Accuracy = 0.8768, F1 Score = 0.4328

Best accuracy and F1 scores for different preprocessing steps

Looking at the preprocessing results, we observe that while the initial model achieved high accuracy (86.46 percent validation), the low F1 score (0.4119) revealed a significant class imbalance issue. Subsequent preprocessing steps, particularly upsampling and polynomial feature expansion, improved the model's ability to handle this imbalance, leading to better F1 scores (0.4323 validation) without compromising overall accuracy (86.81 percent validation). The minimal gap between training and validation accuracy, coupled with improved F1 scores, suggests that our final model achieves a good balance between performance and generalization.

VI. Discussion and comments

Our results show that ridge regression surpassed logistic regression, achieving an F1 score of 0.4328 on the local validation set and 0.434 on the test set submission, along with accuracies of 86.81% and 88%, respectively. This demonstrates ridge regression's capacity to manage the dataset's complexity and high dimensionality effectively, particularly when used with polynomial feature expansion. Additionally, our optimized threshold function proved instrumental in balancing precision and recall, critical for achieving a high F1 score on an imbalanced dataset. Fine-tuning parameters—lambda for regularization, gamma for learning rate, and the classification threshold—was crucial, as each significantly influenced F1 score and overall model performance. Our best submission on AI Crowd achieved an F1 score of 0.439 (ID: 275034). This resulted from not distinguishing between discrete and continuous values when filling missing data with the mean. Our current submission achieve an F1 score of 0.434 but with a better pre-processing.

VII. Conclusion

Our project demonstrated the effectiveness of machine learning in classifying heart disease risk in a complex medical dataset. Ridge regression achieved a higher F1 score than logistic regression by handling high dimensionality and imbalanced classes. Polynomial feature expansion captured key non-

linear relationships, while regularization and threshold tuning improved generalization and precision.

REFERENCES

[1] W. H. Organization. (2021) Cardiovascular diseases (cvds). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)